Going multivariate in clinical trial studies

Kavelaars, Xynthia

Published in:
Small sample size solutions

Publication date:
2020

Document Version
Publisher's PDF, also known as Version of record

10

GOING MULTIVARIATE IN CLINICAL TRIAL STUDIES

A Bayesian framework for multiple binary outcomes

Xynthia Kavelaars

DEPARTMENT OF METHODOLOGY AND STATISTICS, TILBURG UNIVERSITY, TILBURG, THE NETHERLANDS

Introduction

Clinical trials often compare a new treatment to standard care or a placebo. If the collected data provide sufficient evidence that the new treatment is better than the control treatment, the new treatment is declared superior. Since these superiority decisions ultimately contribute to a decision about treatment adoption, proper error control is crucial to ensure that better treatments are indeed selected. Key to regulating decision errors is collecting sufficient information: A quantity that is often expressed in terms of a minimum number of participants, or required sample size.

Recruiting sufficiently large samples can be challenging, however. This is especially true in an era in which medicine is increasingly personalized (Hamburg & Collins, 2010; Ng, Murray, Levy, & Venter, 2009). *Personalization of medicine* refers to the targeting of treatments at specific patient and/or disease characteristics under the assumption that patients with different (disease) characteristics respond differently to treatments (Goldberger & Buxton, 2013). Since personalization limits the target population of the treatment, inclusion and exclusion criteria for trials become more stringent and the eligible number of participants decreases. This inherently decreases the sample size of studies conducted with the same resources. Consequences of small samples may be substantial: Trials may be left underpowered and decisions about superiority might remain inconclusive.

The problem associated with small sample sizes due to stringent inclusion criteria is illustrated by the CAR-B study (Schimmel, Verhaak, Hanssens, Gehring, & Sitskoorn, 2018). CAR-B aims to improve treatment for cancer patients with 11–20 metastatic brain tumors (i.e. tumors that originate from another site in the body and have spread to the brain). These patients have a life expectancy of one or two months

and are currently treated with whole-brain radiation therapy. However, whole-brain radiation has adverse side effects: The treatment damages brain tissue and results in severe cognitive impairment. Local radiation of the individual tumors (stereotactic surgery) is a promising alternative that spares healthy tissue and prevents cognitive decline without increasing mortality. The protective effect on cognition has been demonstrated in a related population of patients with fewer brain tumors (Chang et al., 2009; Yamamoto et al., 2014). However, investigating whether local radiation reduces side effects in the current target population is difficult: Clinicians are reluctant to prescribe the alternative treatment and not all referred patients are eligible for participation, leaving the researchers unable to recruit the required sample.

To improve decision-making with limited samples, studies such as CAR-B might combine information from multiple outcomes. The current chapter introduces a Bayesian decision-making framework to combine two binary outcomes. Since superiority with two outcomes can be defined in multiple ways, several criteria to evaluate treatments are discussed in the "Decision rules" section. Evaluation of these decision rules requires a statistical analysis procedure that combines the outcomes. The "Data analysis" section outlines such a multivariate approach for Bayesian analysis of binary outcomes. The proposed decision-making strategy is illustrated in the "Computation in practice" section, which introduces an online app to analyze real data (for an online version go to https://utrecht-university.shinyapps.io/multiple_binary_outcomes/ – for the annotated R code go to https://osf.io/am7pr/ – and for potential newer versions go to https://github.com/XynthiaKavelaars). Since trials with limited access to participants aim for the smallest sample possible, the chapter continues with "Sample size considerations" to explain how interim analyses during the trial may improve efficiency compared to traditional sample size estimation before running the trial. The "Concluding remarks" section highlights some extensions of the framework. Throughout the chapter, the comparison of local and whole-brain radiation in the CAR-B study serves as an example with cognitive functioning and quality of life as the outcomes under consideration.

Decision rules

A key element of decision-making is the decision rule: A procedure to decide whether a treatment is considered superior. When dealing with two outcomes, superiority can be defined in several ways (Food and Drug Administration, 2017), such as a favorable effect on:

1. The most important outcome ("Single-outcome rule")
2. Both outcomes ("All rule")
3. Any of the outcomes ("Any rule")
4. The sum of outcomes ("Compensatory rule")

Each of these decision rules weighs the effects of the two outcomes differently. The *Single-outcome rule* evaluates the data from one outcome and ignores the

other outcome in the decision procedure. In the CAR-B study, local radiation would be the treatment of preference if it impairs cognitive functioning less than whole-brain radiation, irrespective of the effects on quality of life. The *All rule* evaluates both outcomes, and requires favorable effects on each of them. Compared to whole-brain radiation, more patients should maintain both cognitive functioning *and* quality of life after local radiation. The *Any rule* requires a beneficial effect on at least one outcome and ignores any result on the other outcome. Local radiation would be considered superior if fewer patients experience cognitive side effects, a lower quality of life, or both. The *Compensatory rule* also requires at least one favorable treatment effect, but the compensatory mechanism poses a restriction on the second outcome. The new treatment may perform better, similarly, or even worse than the control treatment on this outcome, but the rule takes the size of the treatment differences into account to weigh beneficial and adverse effects. A net advantage on the sum of outcomes is required, such that several outcome combinations would result in a preference for local radiation. Superiority is concluded as long as favorable effects on cognitive functioning outweigh unfavorable effects on quality of life or vice versa.

The aforementioned decision rules ultimately lead to a conclusion about the treatment *difference*: The new treatment is considered superior if the difference between the new and the control treatment is larger than zero according to the decision rule of interest. For each of the decision rules, the corresponding superiority region is plotted in Figure 10.1. These superiority regions graphically represent how the treatment differences on both individual outcomes should be related to result in superiority: If the probability that the treatment difference falls in the marked area is sufficiently large, the treatment would be declared superior.

Selecting a decision rule

The choice for a decision rule should be guided by the researcher's standard for superiority. To illustrate this, consider the following situations (see Figure 10.2 for a graphical representation):

1. Local radiation performs better on cognitive functioning as well as quality of life
2. Local radiation performs better on cognitive functioning and similarly on quality of life
3. Local radiation performs much better on cognitive functioning and slightly worse on quality of life
4. Local radiation performs slightly better on cognitive functioning and much worse on quality of life

If outcomes are equally important, most researchers would either (a) set a high standard and consider local radiation superior if both outcomes demonstrate an advantage (situation 1), or (b) balance outcomes and consider local radiation superior if advantages outweigh disadvantages (situations 1–3). Situation 4 is

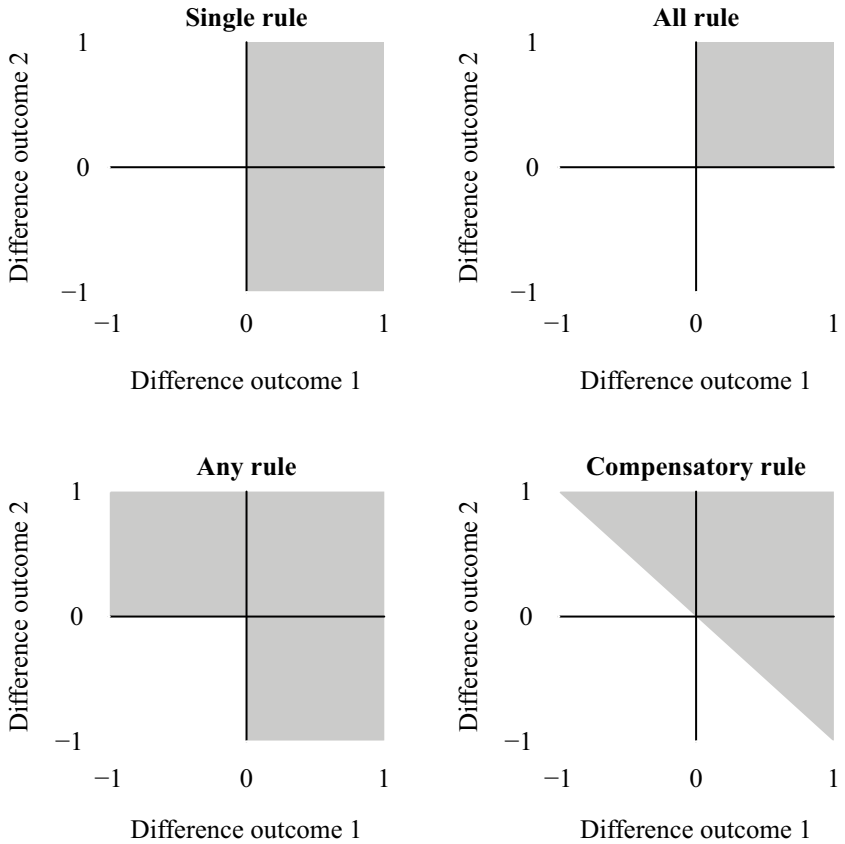

FIGURE 10.1 Superiority regions (shaded areas) for different decision rules

unlikely to result in a preference for local radiation, unless cognitive functioning is much more important than quality of life.

While the All rule applies to the high standard and differentiates situation 1 (superior) from situations 2–4 (not superior), the Compensatory rule balances results and distinguishes situations 1–3 (superior) from situation 4 (not superior). The Single and Any rules do not meet these standards and would conclude that local radiation performs better in all situations, including the fourth. These rules should be used only when unfavorable effects can safely be ignored in the presence of a specific (Single rule) or any (Any rule) favorable effect.

Data analysis

To evaluate the decision rules discussed in the previous section, treatment comparison requires a procedure to quantify evidence in favor of the new treatment. The current section introduces the elements of a Bayesian approach to analyze data from two binary outcomes: likelihood, prior, and posterior distributions.

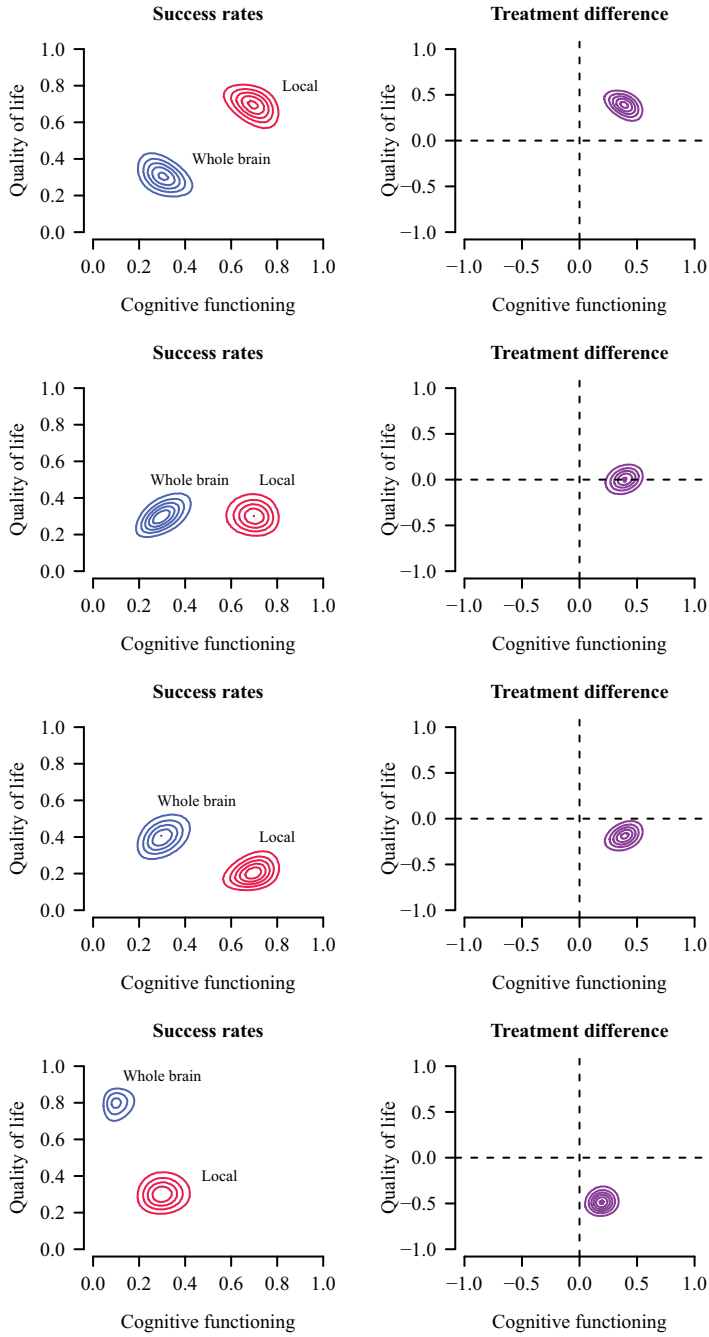

FIGURE 10.2 Example posterior distributions (left panels) and distributions of the treatment difference (right panels) for four different potential treatment differences (local radiation–whole-brain radiation) in the CAR-B study

Description of the data and specification of the likelihood

Binary data have two values, traditionally labeled as 1 for success and 0 for failure. In general, *success* refers to improvement or absence of decline, and *failure* indicates the opposite: decline or absence of improvement respectively. Considering two outcomes together results in two binary responses per participant that can take four different combinations (see Table 10.1). The patient can have successes on both outcomes (x_{11}^{obs}); a success on one outcome, but not on the other (x_{10}^{obs} or x_{01}^{obs}); or failures on both outcomes (x_{00}^{obs}). The total number of successes on a particular outcome equals the sum of simultaneous and separate successes on that outcome, such that $x_1^{obs} = x_{11}^{obs} + x_{10}^{obs}$, etc.

The multivariate likelihood of the outcomes is based on the four response frequencies. These four response frequencies reflect (a) the individual success rates, and (b) the relation between outcomes. The latter serves as an additional source of information that may contribute to more efficient decision-making (Food and Drug Administration, 2010).

Specification of prior information

Prior information represents prior beliefs about success rates of individual treatments as well as the difference between treatments. These prior beliefs can, for example, incorporate information from comparable studies into the current one. Prior beliefs about two binary outcomes are quantified by four prior frequencies, expressed as $x_{11}^{prior}, x_{10}^{prior}, x_{01}^{prior}$, and x_{00}^{prior} (Olkin & Trikalinos, 2015). Each of these individual prior frequencies incorporates information about one of the response frequencies in the data ($x_{11}^{obs}, x_{10}^{obs}, x_{01}^{obs}$ and x_{00}^{obs}). Conveniently, one can think of these prior observations as an extra dataset, where the total number of observations in this prior dataset reflects the strength of the prior beliefs. Strong prior beliefs are translated to many prior observations, whereas weak prior beliefs can be expressed through small numbers of prior observations. An uninformative prior specification for the analysis of two binary outcomes would be a half observation for each response combination, such that the total number of prior observations equals two (Berger, Bernardo, & Sun, 2015). This specification is also called *Jeffrey's prior* and conveys virtually no

TABLE 10.1 Response combinations for two binary outcomes

	Outcome 2		
Outcome 1	*Success*	*Failure*	*Total*
Success	x_{11}	x_{10}	x_1
Failure	x_{01}	x_{00}	$n - x_1$
Total	x_2	$n - x_2$	n

information about the success rates of individual outcomes or the correlation between outcomes. If both treatments have this specification, no prior information about the treatment difference is provided either.

The posterior distribution

The posterior distribution reflects prior beliefs after they have been updated with the data and indicate the posterior success rates of individual outcomes in relation to each other; see also Chapters 1–3 (Miočević, Levy, & Van de Schoot; Miočević, Levy, & Savord; Van de Schoot, Veen, Smeets, Winter, & Depaoli). The posterior response frequencies equal the sum of prior and observed frequencies, such that $x_{11}^{post} = x_{11}^{prior} + x_{11}^{obs}$, etc. Examples of posterior distributions for treatment effects with two outcomes are graphically presented in Figure 10.2.

Comparison of the two posterior distributions allows for decision-making about treatment superiority, by quantifying evidence for a relevant treatment difference as a posterior probability. This posterior probability depends on the definition of superiority as defined via the decision rule and allows for two decisions. If the posterior probability exceeds a pre-specified threshold (often .95 or .99 in clinical trials; Food and Drug Administration, 2010), evidence is strong enough to consider the treatment superior. If the posterior probability is lower than the threshold, there is not sufficient evidence to conclude superiority.

Computation in practice

The online supplement offers a Shiny app to analyze real data using the framework proposed in the previous sections. If the researcher enters the prior $(x_{11}^{prior}, x_{10}^{prior}, x_{01}^{prior}, x_{00}^{prior})$ and observed $(x_{11}^{obs}, x_{10}^{obs}, x_{01}^{obs}, x_{00}^{obs})$ response frequencies for two treatments, the application:

a. Computes the posterior probability of a treatment difference given the introduced decision rules
b. Plots the posterior treatment distributions
c. Plots the posterior distribution of the treatment difference
d. Computes the prior, observed and posterior correlations between outcomes

The Shiny app including user guide can be found at https://utrecht-university.shinyapps.io/multiple_binary_outcomes/ (for the annotated `R` code and potential newer versions go to https://github.com/XynthiaKavelaars).

The method and app are illustrated with artificial data from two treatment distributions with two negatively correlated binary outcome variables ($n = 100$ cases per treatment). The true success probabilities of the experimental and

control treatments were .60 and .40 on both outcomes respectively, such that the experimental treatment performs better on both individual outcomes. The data were used to quantify evidence in favor of the experimental treatment according to the different decision rules (Single, Any, All, Compensatory). The observed response frequencies were entered in the four upper-left cells under "Experimental treatment" and "Control treatment" in the *Data* tab (see Figure 10.3). The app subsequently computed the total observed successes and failures in the margins as well as the observed correlations.

Without any prior knowledge about the treatments or treatment differences, Jeffrey's prior served as a prior distribution, such that each response category was assigned a half observation. After entering the prior frequencies in the *Prior* tab, the app provided the successes and failures per outcome and the prior correlation between outcomes (Figure 10.4).

The *Treatment distributions* tab showed the posterior treatment distributions and posterior correlations of both treatments (Figure 10.5).

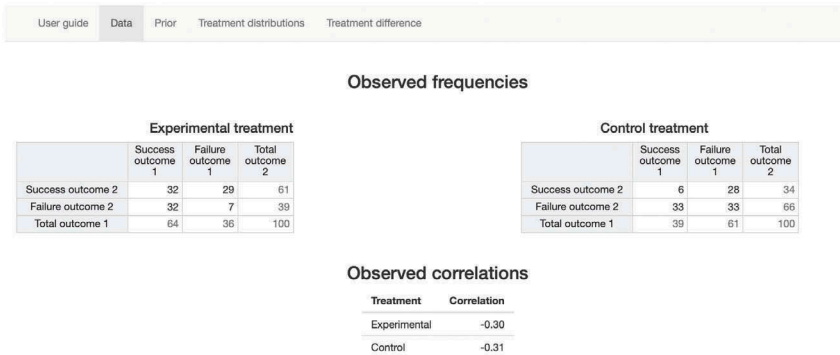

FIGURE 10.3 Screenshot of *Data* tab

FIGURE 10.4 Screenshot of *Prior* tab

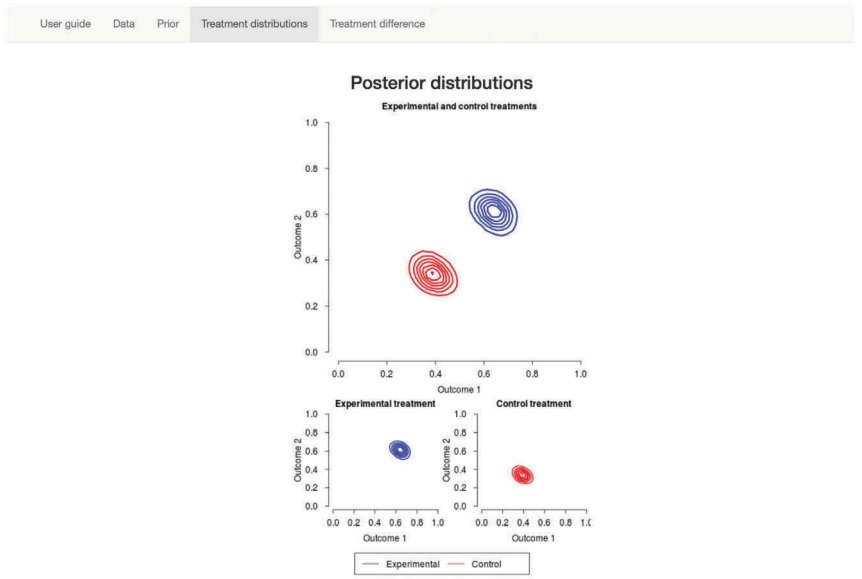

FIGURE 10.5 Screenshot of *Treatment distributions* tab

The *Treatment difference* tab (Figure 10.6) presented the distribution of the posterior treatment difference and the evidence in favor of the experimental treatment according to the proposed decision rules.

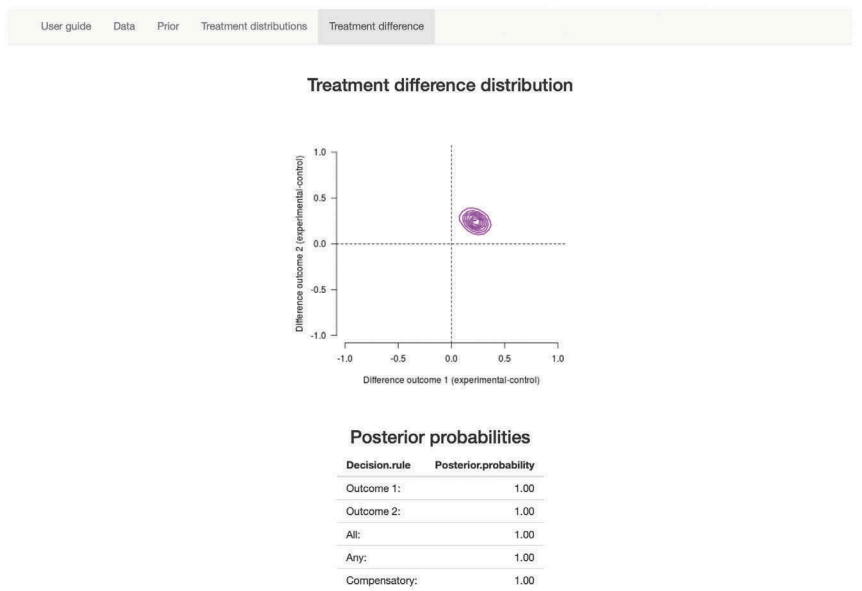

FIGURE 10.6 Screenshot of *Treatment difference* tab

Sample size considerations

When the availability of participants is limited, a highly relevant question is how much data are minimally needed to make a sufficiently powerful decision. Since the sample size traditionally determines when to stop data collection, researchers often estimate the required number of participants before running the trial. Efficient a priori sample size estimation is difficult due to uncertainty about one or multiple treatment differences, regardless of the number of outcomes, since treatment differences are unknown in advance and need to be estimated. However, small inaccuracies in their estimation may have important consequences. Overestimating a treatment difference results in too small a sample to make a powerful decision, while (limited) underestimation needlessly extends the trial.

In trials with multiple outcomes, the required sample size also depends on the decision rule as illustrated in Figure 10.7. The figure shows how evidence in favor of the decision rule under consideration changes for the example data from the "Computation in practice" section, while increasing the sample size in steps of one observation per group. Although the posterior probabilities of all decision rules ultimately approach one and conclude superiority as the data accumulate, different decision rules require different numbers of observations to arrive at that conclusion. With the data presented in Figure 10.7, the Any rule

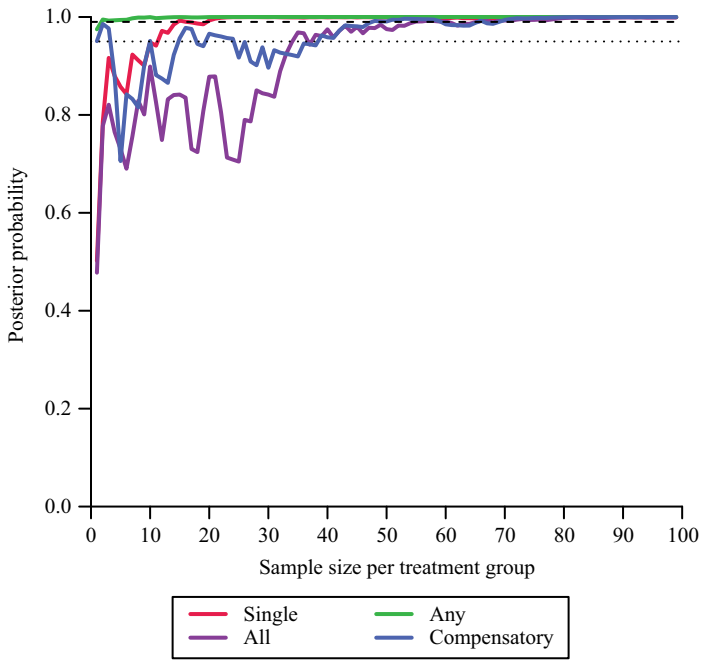

FIGURE 10.7 Example of evidence collection as data accumulate for different decision rules and two different decision criteria (dots = .95; dashes = .99)

requires fewest observations to cross decision thresholds, followed by the Compensatory and Single outcome rules. The All rule requires the largest sample.

The relative efficiency of decision rules displayed in Figure 10.7 is specific to the particular scenario, since different relations between outcomes require different sample sizes to evaluate a specific decision rule (Food and Drug Administration, 2010). To provide an idea of the influence of the correlation between the outcomes, posterior treatment distributions for three correlation structures are displayed in Figure 10.8. This influence affects the proportion of overlap between the distribution of the posterior treatment difference and the superiority region of a decision rule, such that evidence in favor of the new treatment (i.e. posterior probability) as well as the required sample size to reach the decision threshold differ.

Figure 10.9 illustrates how the amount of evidence for each decision rule depends on the correlation when treatment differences are identical. The Single rule is not sensitive to the correlation: The proportion of the difference distribution that overlaps with the superiority region is similar for each correlation structure. The required sample size to conclude superiority will be the same. The All rule has a (slightly) larger proportion of overlap between the distribution of the difference and the superiority region when the correlation is positive. Compared to negatively correlated outcomes, the same amount of evidence can thus be obtained with a smaller sample. The Any and Compensatory rules demonstrate the relationship between the correlation structure and sample size more clearly. The distribution of the treatment difference falls completely in the superiority region when outcomes are negatively correlated (implying a posterior probability of one), while uncorrelated or positively correlated data result in a part of the distribution outside the superiority region (i.e. a posterior probability below one). The sample size will be smallest with negatively correlated outcomes.

In summary, several sources of uncertainty complicate a priori sample size estimation in trials with multiple outcomes: Treatment differences on individual outcomes, the correlation between outcomes, and the decision rule influence the required number of observations. The difficulty of accurately estimating the sample size interferes with the potential efficiency gain of multiple outcomes, such that a priori sample size estimation may be inadequate with small samples and multiple outcomes (Rauch & Kieser, 2015).

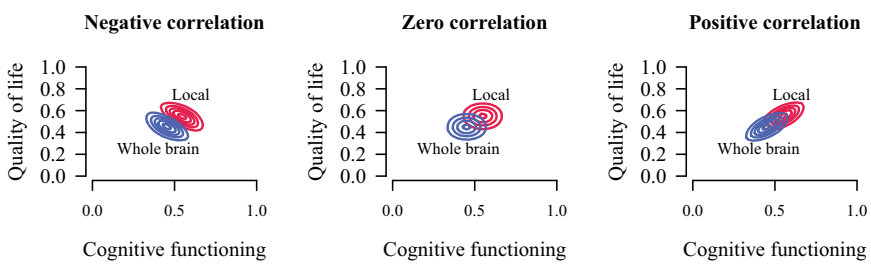

FIGURE 10.8 The influence of the correlation between outcomes on posterior treatment distributions

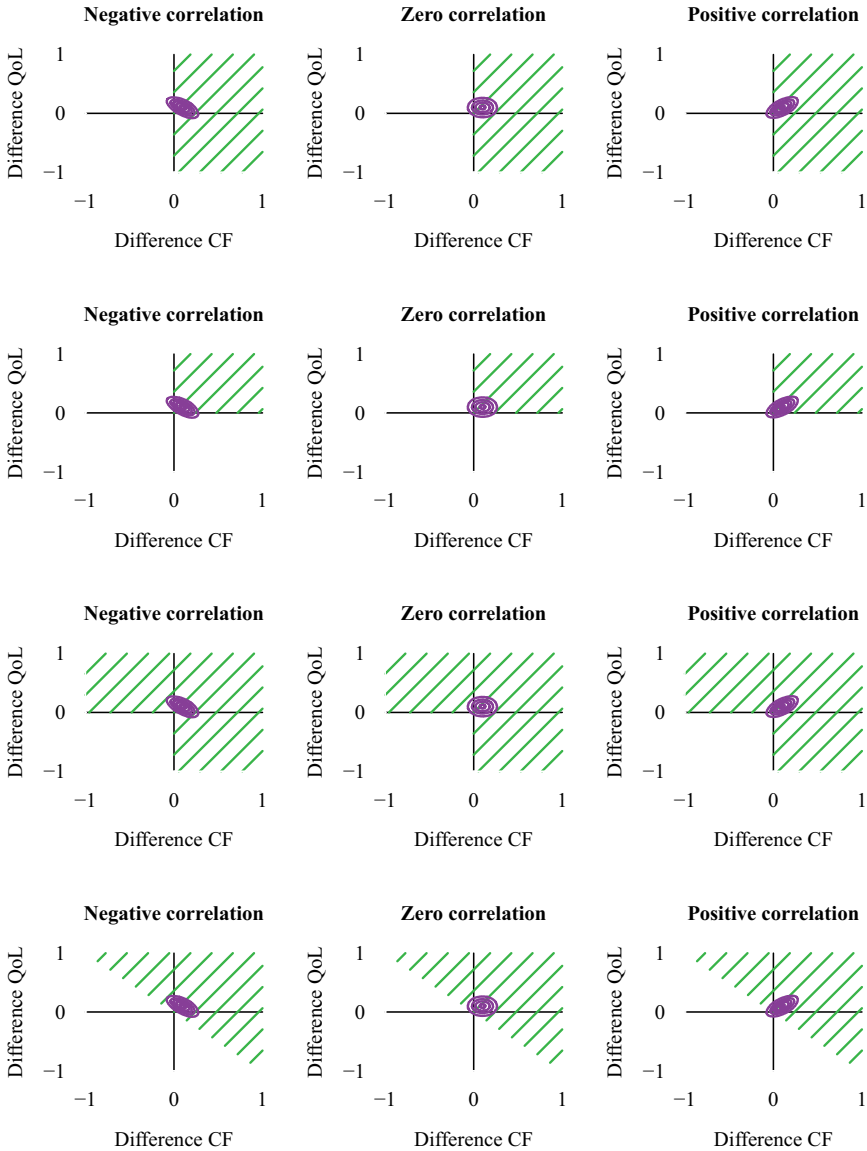

FIGURE 10.9 The influence of the correlation on the evidence for various decision rules. A larger proportion of overlap between the distribution of the treatment difference and the superiority region (shaded area) indicates more evidence. CF = cognitive functioning; QoL = Quality of Life

Adaptive trial design

To reduce the impact of unknown information on the efficiency of trials the sample size can be estimated while running the trial, using a method called

adaptive stopping (Berry, Carlin, Lee, & Muller, 2010). Adaptive stopping performs one or multiple interim analyses and stops the trial as soon as evidence is conclusive, such that efficiency is optimized. Compared to a priori sample size estimation, adaptive stopping may result in early trial termination if the treatment difference is larger than expected (i.e. underestimated). If the treatment difference appears smaller than anticipated (i.e. overestimated) and evidence remains inconclusive, the trial may be extended beyond the planned sample size. Adaptive stopping thus forms a flexible alternative that embraces the uncertainties of the traditional a priori estimated sample size (Bauer, Bretz, Dragalin, König, & Wassmer, 2016; Thorlund, Haggstrom, Park, & Mills, 2018).

Although interim analyses form an attractive approach to improve efficiency, adaptive trials must be designed carefully (Food and Drug Administration, 2010; Sanborn & Hills, 2014). The final decision about superiority potentially requires several interim decisions to evaluate whether evidence is strong enough to draw a conclusion. Without properly adjusting the design to repeated decision-making, the risk of falsely concluding superiority (i.e. Type I error) over all decisions is larger than anticipated, as shown in Figure 10.10 (Sanborn & Hills, 2014). To keep the Type I error risk over *all* decisions acceptable, the Type I error rate for *individual* decisions must be adjusted (Jennison & Turnbull, 1999). A 5% Type I error risk over multiple decisions consequentially results in individual decisions that have a Type I error risk below 5%. The size of the adjustment depends on the number of interim decisions: More decisions require a larger adjustment of the Type I error rate for individual decisions (see Figure 10.10).

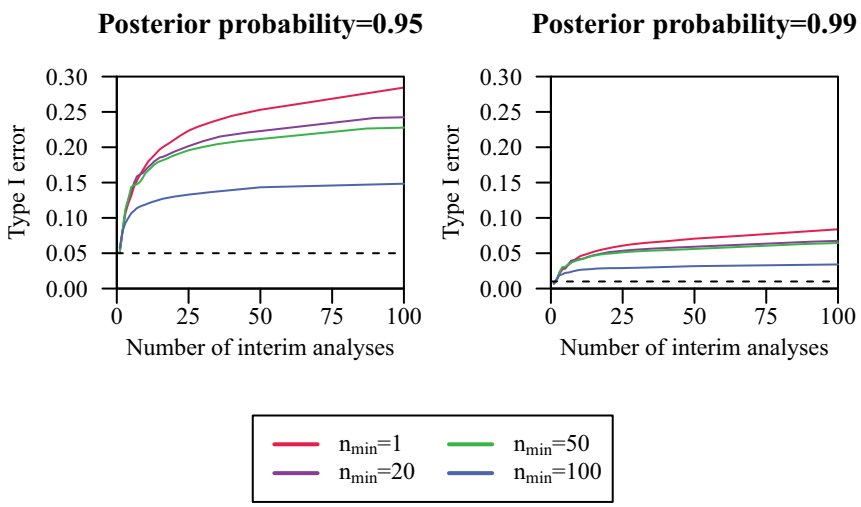

FIGURE 10.10 The empirical Type I error probability as a function of the number of interim analyses for different n_{min} when the decision threshold is not corrected for the number of interim analyses. Dashed lines indicate the desired thresholds of $\alpha = .05$ (posterior probability = .95) and $\alpha = .01$ (posterior probability = .99)

A key element in Type I error control is the decision threshold: the lower limit for the posterior probability to conclude superiority. The decision threshold equals $1-\alpha$, where α is the maximum Type I error probability (Marsman & Wagenmakers, 2017). A 5% risk of an incorrect superiority decision (α = .05) results in a minimal posterior probability of .95. A very high threshold might be attractive to minimize Type I errors, but does not contribute to efficient decision-making: A larger sample size is required to regulate the chance to detect a true treatment difference (i.e. to protect power). The decision threshold thus relates the Type I error and required sample size via the number of interim analyses (Shi & Yin, 2019). Limiting the number of decisions is key to efficiently designing an adaptive trial (Jennison & Turnbull, 1999). To this end, the Food and Drug Administration (2010) recommends balancing the number of interim analyses with decision error rates, by carefully choosing three design parameters:

1. The sample size to look at the data for the first time (n_{min})
2. The number of added participants if the previous analysis did not provide sufficient evidence (interim group size)
3. The sample size to stop the trial if evidence is not strong enough to conclude superiority (n_{max})

The sample size at the first interim analysis (n_{min}) should not be too small for two reasons. First, a small interim sample size could detect unrealistically large treatment effects only and needlessly increases the number of interim analyses. Second, very small samples increase the probability of falsely concluding superiority (Schönbrodt, Wagenmakers, Zehetleitner, & Perugini, 2017). As shown in Figure 10.7, the posterior probability is unstable with few observations and becomes more stable as the number of observations increases. Single observations can be influential in small samples, and this influence diminishes as the sample size increases. A larger n_{min} automatically reduces the number of interim analyses as well as the Type I errors and requires a smaller correction of the decision threshold, as illustrated in Figure 10.10. However, a too large n_{min} limits efficiency: Superiority may have been concluded with a smaller sample and potential participant recruitment is needlessly extended.

If the first interim analysis did not result in conclusive evidence, the sample size can be increased in several steps. The interim group size of added participants should be chosen with the inconclusive results of the previous analysis in mind, such that the new sample provides a reasonable chance of detecting a treatment difference given the earlier lack of evidence. The number of observations between interim analyses may be the same throughout the trial, or can differ per interim analysis if that would benefit the trial's efficiency. It should be chosen carefully, however, since too small and too large group sizes both reduce efficiency (Jennison & Turnbull, 1999). A too small group size needlessly increases the number of interim analyses, while a too large group size reduces the flexibility to terminate the trial as soon as the decision criterion has been met.

Ideally, the sample size to terminate the trial if the data do *not* provide sufficient evidence for superiority (n_{max}) equals the sample size that is required to detect the smallest treatment effect of clinical interest (Food and Drug Administration, 2010). In practice, n_{max} will often be limited by the maximum number of available participants and may be smaller than optimal, which has the same consequence as a too small (a priori estimated) sample size: A limited n_{max} restricts the power to detect small treatment differences.

Concluding remarks

The current chapter presented a Bayesian framework for decision-making with multiple outcomes and illustrated how decisions with two outcomes may help a small sample, when (a) using a decision rule that combines information from two outcomes efficiently, and (b) designing a trial adaptively. Without giving all the mathematical details, I have tried to provide a clear intuition to the approach and software to carry out the analysis.

The proposed approach has several extensions that may accommodate more realistic decisions. First, more than two outcomes can be included, such that researchers might weigh treatment differences on three or more relevant aspects. Increasing the number of outcomes may further improve efficiency, but more outcomes also increase the complexity of the data analysis.

Second, although equal importance of outcomes was assumed throughout the chapter, unequal importance of outcomes could be incorporated. The Compensatory rule in particular could be adapted easily to, for example, include survival into a decision; an outcome that is in many cases more important than cognitive side effects. However, user-friendly software packages for more outcomes remain to be developed.

Third, the applicability of adaptive designs can be strongly improved with clear guidelines on the concrete choice of design parameters. Optimal design of interim analyses is necessary to do justice to the potential flexibility of adaptive trials.

Acknowledgement

This work was supported by a NWO (Dutch Research Council) research talent grant (no. 406.18.505). I thank Maurits Kaptein and Joris Mulder (both Tilburg University, The Netherlands), and the reviewers for sharing their insights and providing comments that greatly improved the manuscript.

References

Bauer, P., Bretz, F., Dragalin, V., König, F., & Wassmer, G. (2016). Twenty-five years of confirmatory adaptive designs: Opportunities and pitfalls. *Statistics in Medicine*, *35*(3), 325–347.

Berger, J. O., Bernardo, J. M., & Sun, D. (2015). Overall objective priors. *Bayesian Analysis*, *10*(1), 189–221.

Berry, S. M., Carlin, B. P., Lee, J. J., & Muller, P. (2010). *Bayesian adaptive methods for clinical trials*. Boca Raton, FL: Chapman & Hall/CRC Press.

Chang, E. L., Wefel, J. S., Hess, K. R., Allen, P. K., Lang, F. F., Kornguth, D. G., Meyers, C. A. (2009). Neurocognition in patients with brain metastases treated with radiosurgery or radiosurgery plus whole-brain irradiation: A randomised controlled trial. *Lancet Oncology*, *10*(11), 1037–1044.

Food and Drug Administration. (2010). Guidance for industry: Adaptive design clinical trials for drugs and biologics. Center for Drug Evaluation and Research, US Food and Drug Administration, Silver Spring, MD. https://wayback.archive-it.org/7993/20170403220223/https://www.fda.gov/ucm/groups/fdagov-public/@fdagov-drugs-gen/documents/document/ucm201790.pdf.

Food and Drug Administration. (2017). Multiple endpoints in clinical trials guidance for industry. Center for Biologics Evaluation and Research (CBER). www.fda.gov/regulatory-information/search-fda-guidance-documents/multiple-endpoints-clinical-trials-guidance-industry. www.fda.gov/media/102657/download.

Goldberger, J. J., & Buxton, A. E. (2013). Personalized medicine vs guideline-based medicine. *JAMA*, *309*(24), 2559–2560.

Hamburg, M. A., & Collins, F. S. (2010). The path to personalized medicine. *New England Journal of Medicine*, *363*(4), 301–304.

Jennison, C., & Turnbull, B. W. (1999). *Group sequential methods with applications to clinical trials*. New York, NY: Chapman & Hall/CRC Press.

Marsman, M., & Wagenmakers, E.-J. (2017). Three insights from a Bayesian interpretation of the one-sided P value. *Educational and Psychological Measurement*, 77(3), 529–539.

Ng, P. C., Murray, S. S., Levy, S., & Venter, J. C. (2009). An agenda for personalized medicine. *Nature*, *461*(7265), 724–726.

Olkin, I., & Trikalinos, T. A. (2015). Constructions for a bivariate beta distribution. *Statistics & Probability Letters*, *96*, 54–60.

Rauch, G., & Kieser, M. (2015). Adaptive designs for clinical trials with multiple endpoints. *Clinical Investigation*, *5*(5), 433–435.

Sanborn, A. N., & Hills, T. T. (2014). The frequentist implications of optional stopping on Bayesian hypothesis tests. *Psychonomic Bulletin & Review*, *21*(2), 283–300.

Schimmel, W. C. M., Verhaak, E., Hanssens, P. E. J., Gehring, K., & Sitskoorn, M. M. (2018). A randomised trial to compare cognitive outcome after gamma knife radiosurgery versus whole brain radiation therapy in patients with multiple brain metastases: Research protocol CAR-study B. *BMC Cancer*, *18*(1), 218.

Schönbrodt, F. D., Wagenmakers, E.-J., Zehetleitner, M., & Perugini, M. (2017). Sequential hypothesis testing with Bayes factors: Efficiently testing mean differences. *Psychological Methods*, *22*(2), 322.

Shi, H., & Yin, G. (2019). Control of Type I error rates in Bayesian sequential designs. *Bayesian Analysis*, *14*(2), 399–425.

Thorlund, K., Haggstrom, J., Park, J. J. H., & Mills, E. J. (2018). Key design considerations for adaptive clinical trials: A primer for clinicians. *BMJ*, *360*, k698.

Yamamoto, M., Serizawa, T., Shuto, T., Akabane, A., Higuchi, Y., Kawagishi, J., Tsuchiya, K. (2014). Stereotactic radiosurgery for patients with multiple brain metastases (JLGK0901): A multi-institutional prospective observational study. *Lancet Oncology*, *15*(4), 387–395.